软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1451−1476 [doi: 10.13328/j.cnki.jos.006461]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于数据场聚类的共享单车需求预测模型*

乔少杰[1], 韩 楠[2], 岳 昆[3], 易玉根[4], 黄发良[5], 元昌安[6], 丁 鹏[1], Louis Alberto GUTIERREZ[7]

[1](成都信息工程大学 软件工程学院, 四川 成都 610225)

[2](成都信息工程大学 管理学院, 四川 成都 610225)

[3](云南大学 信息学院, 云南 昆明 650504)

[4](江西师范大学 软件学院, 江西 南昌 330022)

[5](南宁师范大学 计算机与信息工程学院, 广西 南宁 530023)

[6](广西教育学院, 广西 南宁 530023)

[7](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

通信作者: 韩楠, E-mail: hannan@cuit.edu.cn

摘 要: 共享单车系统日益普及, 积累了海量的出行轨迹数据. 在共享单车系统中, 用户的借车和还车行为是随机的, 且受天气、时间等动态因素影响, 使得共享单车调度不平衡, 影响单车用户体验, 并给运营商造成巨大经济损失. 提出了新型基于站点聚类的共享单车需求预测算法, 通过构建单车转移网络计算站点活跃度, 充分考虑站点地理位置和单车转移模式因素, 基于数据场聚类思想, 将距离相近和用车模式相似的站点聚合到一个聚簇中, 给出最佳簇中心个数求取方法. 充分分析时间和天气因素对站点单车需求的影响, 利用皮尔逊相关系数, 从真实天气数据中选择相关性最大的天气特征, 结合历史聚簇内单车需求量, 将其转化为三维向量, 利用多特征长短时记忆深度神经网络LSTM (long short-term memory)对向量内的特征信息进行学习和训练, 以30分钟为长时间间隔, 对每个聚簇内的单车需求量进行预测分析. 与传统机器学习算法和当前主流方法进行对比, 实验结果表明, 所提单车需求模型预测性能得到显著提升.

关键词: 共享单车系统; 单车转移网络; 站点聚类; 数据场; LSTM 网络

中图法分类号: TP18

中文引用格式: 乔少杰, 韩楠, 岳昆, 易玉根, 黄发良, 元昌安, 丁鹏, Gutierrez LA. 基于数据场聚类的共享单车需求预测模型. 软件学报, 2022, 33(4): 1451−1476. http://www.jos.org.cn/1000-9825/6461.htm

英文引用格式: Qiao SJ, Han N, Yue K, Yi YG, Huang FL, Yuan CA, Ding P, Gutierrez LA. Shared-bike Demand Prediction Model Based on Station Clustering. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1451−1476 (in Chinese). http://www.jos.org.cn/1000-9825/6461.htm

Shared-bike Demand Prediction Model Based on Station Clustering

QIAO Shao-Jie[1], HAN Nan[2], YUE Kun[3], YI Yu-Gen[4], HUANG Fa-Liang[5], YUAN Chang-An[6], DING Peng[1], Louis Alberto GUTIERREZ[7]

[1](School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

[2](School of Management, Chengdu University of Information Technology, Chengdu 610225, China)

[3](School of Information Science and Engineering, Yunnan University, Kunming 650504, China)

[4](School of Software, Jiangxi Normal University, Nanchang 330022, China)

* 基金项目: 国家自然科学基金(61772091, 61802035, 61962006, 62072311, U1802271, U2001212); 四川省科技计划(2021JDJQ0021, 2020YFG0153, 20YYJC2785, 2019YFS0067, 2020YJ0481, 2020YFS0466, 2020YJ0430, 2020YDR0164); CCF-华为数据库创新研究计划(CCF-HuaweiDBIR2020004A); 广西自然科学基金(2018GXNSFDA138005)

本文由“面向开放场景的鲁棒机器学习”专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.

收稿时间: 2021-01-17; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

[5](School of Computer and Information Engineering, Nanning Normal University 530023, Nanning, China)
[6](Guangxi College of Education, Nanning 530023, China)
[7](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

Abstract: Bike-sharing system is becoming more and more popular and there accumulates a large volume of trajectory data. In the bike-sharing system, the borrowing and returning behavior of users are arbitrary. In addition, bike-sharing system will be affected by weather, time period, and other dynamic factors, which makes shared bike scheduling unbalanced, affects user's experience, and causes huge economic losses to operators. A novel shared-bike demand prediction model based on station clustering is proposed, the activeness of stations is calculated by constructing a bike transformation network. The geographical location of stations and the bike transmission patterns are taken into full consideration, and the stations with near distances and transformation patterns are aggregated into a cluster based on the idea of data field clustering. In addition, a method for computing the optimal number of cluster centers is presented. The influence of time and weather factors on bike demand is fully analyzed and the Pearson correlation coefficient is used to choose the most relevant weather features from the real weather data and transformed into a three-dimensional vector by taking into consideration the historical demand for bicycles in the cluster. In addition, long short-term memory (LSTM) neural network with multiple features is employed to learn and train the feature information in the vector, and the bike demand in each cluster is predicted and analyzed every thirty minutes. When compared with the traditional machine learning algorithms and the state-of-the-art methods, the results show that the prediction performance of the proposed model has been significantly improved.

Key words: bike-sharing system; bike transformation network; station clustering; data field; long short-term memory (LSTM) network

近年来, 共享单车成为一种主要的出行手段, 成为智慧城市中不可或缺的交通工具. 随着环保意识的提升, 越来越多的人更加重视以绿色环保的方式出行. 此外, 共享单车真正解决了"最后一公里"问题, 改变了人们的生活方式. 截至 2020 年 10 月底, 哈啰单车用户累计骑行 240 亿公里, 累计减少碳排放量近 280 万吨[1]. 一定数量的共享单车, 为用户提供点到点的出行方案, 可以有效改善交通拥堵现象.

虽然共享单车为出行带来诸多便利, 成为一种主流出行方式, 但是通过对国内外当前研究现状的分析, 了解到如何以高效的方式运营共享单车系统具有一定挑战性, 现有研究成果的局限性主要归纳为如下几点.

(1) 国内外现有研究很多都是针对单个站点的需求进行预测, 没有考虑站点间的关联对单车使用的影响. 为了解决整个城市内单车供求不平衡问题, 仅研究单个站点的需求量, 不足以提升共享单车系统的服务质量;

(2) 虽然目前已经存在一些单车需求预测模型, 但是这些模型普遍存在区域局限性问题. 通常在一个城市的预测准确度比较高, 但出于某些原因, 如用户出行习惯不同或天气差异较大等, 对其他城市的单车需求预测效果并不理想;

(3) 现有单车需求预测研究根据经验仅考虑了静态条件因素, 忽略了任意时间长度以前的动态因素对单车需求的影响. 在未来一段时间内, 用户对单车的需求量会受当前以及前一段时间单车站点状态的影响.

本文的研究动机基于如下几点考虑: 1) 现有工作没有考虑天气因素对站点内单车需求的影响, 仅利用历史单车行程记录和站点分布进行分析. 通过前期大量实验发现: 在共享单车系统中, 天气信息是影响需求量的重要因素, 考虑天气特征可以极大地提高算法的准确性; 2) 对不同时间段站点内单车使用情况进行分析, 发现大多数站点的单车使用模式呈现多样性, 一定范围的站点内单车使用模式更具相似性. 然而现有研究主要是对单一站点内单车需求进行预测, 准确性相对较低; 3) 通过深度学习的方法可以考虑任意时间长度以前的动态因素对单车需求量的影响, 而传统的单车需求预测算法根据经验仅考虑了静态条件因素, 因此无法满足站点内单车的实时需求预测.

为了克服现有单车需求预测方法的不足, 本文提供新型基于数据场聚类和长短时记忆深度神经网络的智能共享单车需求预测模型, 主要贡献包括: (1) 构建单车转移网络, 分别考虑借车站点对自身活跃度的影响和还车站点对借车站点活跃度的影响, 得到所有站点的活跃度; (2) 综合考虑站点位置和单车转移模式, 利用数据场聚类的思想, 对共享单车系统内的站点进行二级聚类, 利用轮廓系数求取最佳的簇中心个数; (3) 分析时间和天气因素对站点单车需求的影响, 进而选择影响预测准确性的关键特征; (4) 构建三维向量, 利用多特征

LSTM 网络长时间记忆的特性分析天气和共享单车数据, 更加准确地预测聚簇在不同时间段和不同天气状况下的单车需求.

本文第 1 节综述流行的单车预测模型. 第 2 节给出本文所提方法使用的主要概念. 第 3 节详细分析影响单车需求的天气和时间特征. 第 4 节介绍基于数据场的二级站点聚类算法. 第 5 节给出基于 LSTM 的共享单车预测模型并分析算法复杂度. 第 6 节对预测模型的实验结果进行分析. 第 7 节对本文工作进行总结并探讨未来工作.

1 相关工作

人们活动的随机性不可避免地会造成站点内可用单车数量和停车桩数量无法满足实时需求, 即用户在某一时刻无车可借和无桩可停. 各个站点所处位置和单车使用规律存在差异, 站点需求预测问题是指通过分析各个站点借车和还车的规律来预测未来一段时间内各个站点的需求量, 进而提高单车系统的服务质量. 单车需求预测算法大致分为两类: 基于 station-level 的算法和基于 cluster-level 的算法, 代表性工作如下.

(1) 基于 station-level 的算法

此类算法将每个站点作为研究个体, 分别预测各个站点内单车的需求量. Yang 等人[2]结合考虑站点间单车历史使用数据和天气数据把单车系统模拟为一个动态网络, 提出了基于蒙特卡洛的预测算法. 该算法建立了一个移动模型描述站点间单车的时空转换特点, 基于线上借车记录估计用户在某一站点还车的概率. Huang 等人[3]针对每个站点共享单车数量预测问题提出了基于高斯泊松分布的算法, 基本思想是: 使用非齐次泊松描述借车和还车过程, 近似估计外部因素对单车使用的影响, 计算借还车数量的平均值估计单车的使用量. 算法考虑了外部因素对单车使用的影响, 不足之处在于训练时间较长. Ashqar 等人[4]使用随机森林和最小二乘提升法对每个站点进行预测, 使用最小二乘回归算法来减少共享单车网络中站点预测模型的数量. 此外, 利用随机森林和最小二乘提升法研究多种因素对预测结果的影响. 实验表明: 用随机森林效果比最小二乘提升算法好, 但有些站点的预测误差仍然较大. Liu 等人[5]根据历史行程记录建立模型预测站点取车数量, 对站点与站点之间的联系和行程持续时间进行分析, 提出了预测站点放车数量的模型. 将天气各个指标进行分级处理, 使用高斯核函数计算天气环境相似度, 给每个指标赋予一个权重, 利用 top-k 进行取车数量预测. 针对某一时间段内某一站点放车的数量预测问题, 通过分析典型 station-station 单车持续时间的分布来估计两个站点的骑车时间, 进而实现放车数量的估计. Fricker 等人[6]提出了一种随机模型, 用于研究用户随机选择站点的行为对站点内单车需求不平衡的影响, 通过量化站点内单车的使用模式, 进而最大限度地减少不平衡站点的数量.

(2) 基于 cluster-level 的算法

该类算法认为预测每个站点内共享单车的需求比较困难, 因为每个站点的单车使用模式是动态的且易受各种因素[7,8], 如时间、天气、随机事件等的影响. 因为站点间存在某些相似特征, 基于 cluster-level 的算法将相似的站点划分到同一个簇中, 以簇作为研究对象, 分别预测簇内单车的需求量. Li 等人[9]提出一种层次型预测模型来预测一段时间后簇内每个站点取车和放车的数量, 所提方法的优势在于: 可以处理不平衡天气分布问题; 保证簇间取车概率总和为 1. 不足之处在于: 所提的层次预测模型仅将站点聚合成静态簇, 没有考虑外部环境的影响; 且当外部环境变化时, 无法保证算法准确性. 针对上述不足, Chen 等人[10]将影响单车使用量的因素分为一般和特殊两类, 提出了基于聚类的动态预测算法, 根据当前环境构建一个权重关系网络来模拟单车站点间的关系, 将具有相似用车模式的邻近站点动态划分到一个簇内. 该算法综合考虑了多种影响单车使用量的因素, 预测效果相比其他预测单个站点单车使用量的算法具有更高准确性. Feng 等人[11]认为距离较近的几个站点相较于单个站点的使用模式更具有规律性, 提出了层次流量预测模型来预测每个簇内单车的取车和放车数量, 使用迭代谱聚类算法对站点进行划分, 结合标签传播算法[12]控制簇的地理范围, 通过梯度上升回归树预测整个单车系统的取车总量, 基于预测比例推导每个簇内取车量. 此外, 提出了簇内转移比例模型, 描述簇间取车和还车的关系, 进而预测簇内放车数量. Schuijbroek 等人[13]通过分析共享单车系统的主要

运营成本, 结合站点服务水平要求和调度单车的最佳路由两个因素, 同时考虑服务可行性与近似路由的成本, 提出了先聚类后路由的启发式算法.

对于输入的大规模数据, 卷积神经网络无需人为干预, 可以自动进行特征提取, 被大量应用于遥感科学、计算机视觉、自然语言处理等方面. 例如: 在计算机视觉方面, 图像识别常常使用卷积神经网络来提取图像特征供分类器学习. 此外, 卷积神经网络被应用于无人驾驶中的物体识别. Lin 等人[14]提出一种新的基于数据驱动图滤波的图卷积神经网络模型, 构建了 4 个典型的数据矩阵, 即空间距离矩阵、需求矩阵、平均行程时间矩阵和需求关联矩阵, 量化站点间隐藏的异构关系, 估计图谱结构中的参数, 进而预测站点内每个小时的单车需求量. Chai 等人[15]提出一种多图卷积神经网络模型, 每个图中的结点表示单车站点, 边表示站点间的关系, 构建了多个图进而挖掘站点间的距离和历史用车关联, 对图进行融合并应用于卷积层, 进而预测每个站点的取车和放车数量. 但是这一方法没有考虑天气因素对预测结果的影响, 导致准确率不能得到保证. Lü 等人[16]使用大规模的交通数据, 提出一种基于深度结构模型的交通流量预测算法, 该算法考虑了空间和时间的内在关联, 利用叠层自动编码器模型学习一般的交通流量特征, 并以贪婪的分层方式进行训练.

长短时记忆网络解决了传统神经网络存在的梯度消失问题, 并且具有长时间记忆的特性, 在语音文本和时序预测等研究领域得到广泛应用. Xu 等人[17]提出了一种基于长短期记忆网络的温度预测算法, 使用数据中心温度监控数据和服务器的实际运行参数, 生成时间序列训练集来训练神经网络模型, 进而预测服务器入口温度. Su 等人[18]利用 LSTM 网络进行情绪障碍检测, 通过对情绪障碍引起反应的时间信息建模, 提高传统监检测算法的准确性. 本文利用 LSTM 网络具有长时间记忆的特性, 将其应用于共享单车时序模型预测, 进而计算单车的需求.

综上, 为了克服传统单车需求预测算法的不足, 包括仅考虑历史用车记录导致预测准确率不高、训练时间长、人工设置的参数值较多等问题, 本文提出了一种基于站点聚类和深度学习的共享单车需求预测模型, 通过大量实验证明: 所提模型在提高共享单车需求预测准确性的同时, 可以保证算法的实效性.

2 基本概念

本节主要对算法中使用的概念进行形式化定义, 如下所示.

定义 1(单车转移网络). 单车转移网络定义为一个加权有向图网络:

$$G=(D,T),\ D=\{\sum d_i \mid i=1,2,...,n\},\ T=\{\sum t_{ij} \mid i,j=1,2,...,n\},$$

其中, D 表示单车站点的集合, T 表示边的集合, t_{ij} 表示从站点 d_i 到站点 d_j 的转移边.

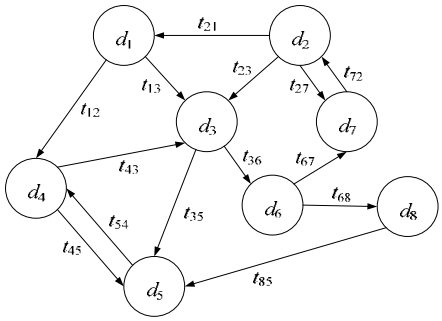
如图 1 所示: 在单车转移网络中, 站点间存在转移边, 每条边包含一个属性 t_{ij}, 代表从站点 d_i 到站点 d_j 的单车转移量 n.

定义 2(活跃度贡献率). 活跃度贡献率 p_{ij} 代表站点 d_j 对站点 d_i 活跃度大小的影响. 如果站点 d_j 内的单车数量为 n_j, 其中, 从站点 d_i 转移到站点 d_j 的单车数量为 $n_{i,j}$, 则 $n_{i,j}/n_j$ 表示站点 d_j 对站点 d_i 活跃度贡献率 p_{ij}.

定义 3(站点活跃度). 站点活跃度 v_i 表示站点 d_i 被用户访问的概率, 用于衡量站点 d_i 的重要性. 站点 d_i 的活跃度计算方法见公式(1):

$$v_i=(1-\xi)\frac{n_i^o}{\sum_{k=1}^{l} n_k^o}+\xi\sum j=O(i)\,p_{ij}v_j \tag{1}$$

其中, ξ表示阻尼因子; n_i^o 表示从站点 d_i 借出的单车数量; $O(i)$表示从站点 d_i 借车的其他站点的集合; p_{ij} 为活跃度贡献率, 表示站点 d_j 对站点 d_i 活跃度大小的影响程度.



单车站点集合: $\langle d_1,d_2,d_3,d_4,d_5,d_6,d_7,d_8\rangle$
边集合: {⟨1,2⟩,⟨1,3⟩,⟨2,1⟩,⟨2,3⟩,⟨2,7⟩,⟨3,5⟩,⟨3,6⟩,⟨4,3⟩,⟨4,5⟩,⟨5,4⟩,⟨6,7⟩,⟨6,8⟩,⟨7,2⟩,⟨8,5⟩}

图 1 共享单车转移网络拓扑图

定义 4(单车转移矩阵). 单车转移矩阵 M_i 表示站点 d_i 内单车转移模式, 即站点 d_i 内单车在特定时间段转移到其他站点聚簇的概率.

例 1(单车转移矩阵示例): 公式(2)表示站点 d_1 的单车转移模式, 矩阵的行对依次应 7 个时间段: 工作日的 7:00–11:00(早高峰), 11:00–17:00(上班), 17:00–21:00(晚高峰), 21:00–7:00(休息)和节假日的 0:00–9:00(休息), 9:00–19:00(休闲)和 19:00–24:00(休息), 矩阵的列分别表示 3 个站点聚簇. 矩阵元素$(M_i)_{k,j}$表示在时间段 k 内从站点 d_i 借的单车停留在聚簇 C_j 的概率, 如: $(M_1)_{2,3}$ 表示在时间段 11:00–17:00 从站点 d_1 内借的单车停在聚簇 C_3 的概率为 0.6.

$$M_1=\begin{pmatrix}0.2 & 0.3 & 0.5\\ 0.3 & 0.1 & 0.6\\ 0.4 & 0.4 & 0.2\\ 0.1 & 0.7 & 0.2\\ 0.1 & 0.1 & 0.8\\ 0.5 & 0.2 & 0.3\\ 0.3 & 0.3 & 0.4\end{pmatrix} \tag{2}$$

定义 5(单车需求量). 借车需求量 $B_{i,t}$ 代表 t 时间段内在聚簇 C_i 中借车的数量. 相应地, 还车需求量 $R_{i,t}$ 代表在 t 时间段内在聚簇 C_i 还车的数量.

3 特征分析和矩阵表达

在考虑需求量预测问题时, 需要确定与单车需求相关的主客观因素的数据特性, 进而分析这些数据对预测结果的影响. 单车需求量的变化与时间和天气有关, 本文首先分析时间和天气特性, 然后将他们作为关键因素, 预测在不同时刻、不同天气状况下的单车需求量. 天气数据包含许多特征, 因此选择与单车需求相关的特征显得尤为重要. 在本研究中, 通过分析天气和时间数据与单车需求的相关性, 根据各种特征与单车需求的相关系数对数据进行排序, 选择排名靠前的 k 个作为重要特征. 本文使用皮尔逊相关系数[19]作为特征相关性的标准, 定义如下.

定义 6(皮尔逊相关系数).

$$\gamma=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{3}$$

其中, x_i 和 y_i 分别表示第 i 个特征值和对应的用车量, $\overline{x}$ 和 $\overline{y}$ 分别表示第 i 个特征的平均值和用车量的平均值, n 表示数据规模. γ值越接近 1, 表示 x 与 y 相关程度越高, 即特征 x 对用车量 y 影响越大.

本节重点分析影响共享单车需求预测的动态特征因素, 根据特征选择规则选取相关性较大的特征, 完成

模型构建后, 将选取的特征构建为二维动态特征矩阵, 输入到预测模型中, 并利用长短时记忆深度神经网络构建单车需求预测模型.

3.1 时间特征分析

共享单车的使用频率一般会随着时间的变化而产生波动. 用户在工作日的用车习惯更有规律性, 一般在上下班高峰期用单车往来于公交站与地铁站, 所以用车量相对集中. 但在夜间, 随着公交地铁等公共交通下班, 共享单车使用量会出现上下波动. 在双休日, 用户使用单车的习惯具有较大随机性. 此外, 单车的使用量在同一天的不同时刻也会发生变化, 且相邻时刻之间用户对单车的需求存在关联性. 本文分别从星期和时刻两个方面对共享单车系统进行分析.

(1) 工作日和双休日分析.

图 2 显示了 2014 年 6 月 16 日–22 日的一周内, 所有站点的单车借用量分布. 其中, 点状柱条表示用户在工作日的单车借用量, 斜线柱条表示用户在双休日对单车借用量.

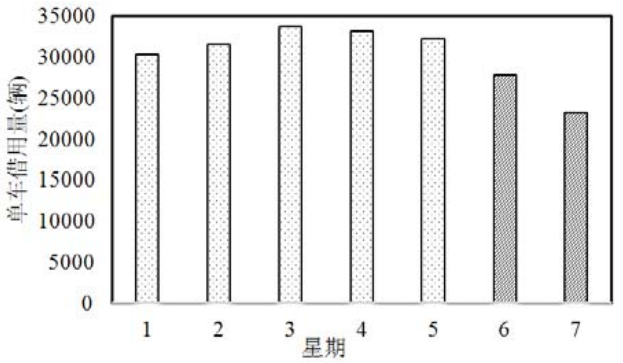


图 2 一周内所有站点的单车借用量分布

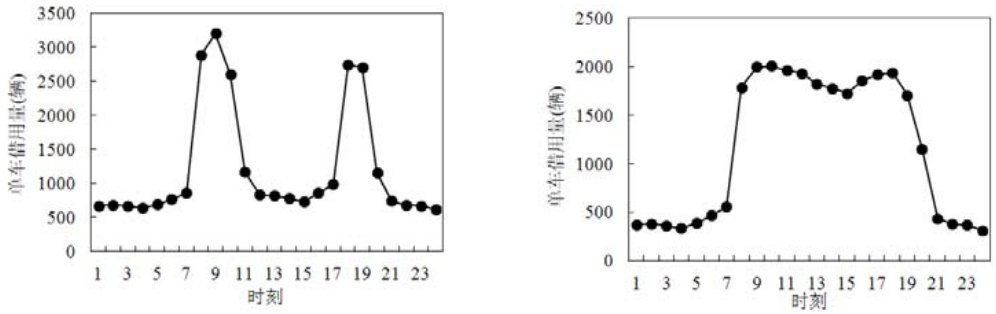
从图中可以发现: 用户在工作日期间对单车的需求较大, 借用量都在 3 万辆以上; 此外, 用户对单车使用量的波动不明显, 稳定在 3 万–3.5 万之间; 但是在双休日期间, 单车的使用量都低于 3 万, 而且周六周日的借用量还出现较大波动; 相对于周六而言, 周日的单车借用量明显减少. 可能的原因是: 用户考虑周一即将到来, 需要在家好好休息, 即使如图 2 所示的周日是多云天气, 也不会对出行造成特殊影响.

(2) 同一天的不同时刻分析.

根据用户在工作日和双休日对单车的需求不同, 对 2014 年 4 月 1 日–5 月 30 日期间的行程记录数据, 按照工作日和双休日分开考虑, 分别分析同一天内不同时刻对单车使用量, 如图 3 所示. 图 3 表明: 上下班高峰期(8:00–10:00 和 18:00–20:00)对单车需求明显增多, 最高达到 3 200 辆, 而在其余时刻的需求量趋于平缓. 而在图 3 中可以发现: 双休日对单车的需求集中在 9:00–19:00 时间段, 且分布均匀, 但是对单车的整体需求还是低于工作日, 最高需求量在 2 000 辆附近. 可以发现: 无论工作日还是双休日, 一天内不同时刻对单车的使用量都会产生影响.

3.2 天气特征分析

天气对共享单车的使用情况具有显著影响, 因为天气会影响到用户的安全出行, 其影响因素包括天气状况、温度、湿度等. 通常, 在晴天单车使用量比下雨天明显增多, 下雨天用车条件恶劣, 比如道路积水易滑等. 当可见度较低时, 用户视野会变窄, 安全性降低, 用户会选择其他方式出行. 此外, 当室外温度适宜、风速较小、湿度低的情况下单车使用率更高, 用户更容易在这种环境条件下出行. 本文分别从天气状况、温度和湿度这 3 个方面对共享单车借用量进行分析.



(a) 工作日 (b) 双休日

图 3 工作日和双休日不同时刻所有站点单车借用量

(1) 天气状况分析

为了分析不同天气状况下共享单车的使用量, 从天气数据中选取了 5 个对单车使用影响最大的天气条件, 分别为晴天、多云、阴天、小雨和大雨. 图 4 显示了在 5 种天气条件下, 每小时内所有站点的单车借用量. 可以发现: 正常情况下, 单车的使用量都很高, 在晴天、多云和阴天的出行条件, 每小时单车借用量保持在 1 300 辆以上; 而在恶劣的天气条件下(下雨), 单车的使用量会受到极大影响, 每小时内单车的使用量急剧下降, 且随着恶劣的程度有更严重的下降趋势.

(2) 温度分析

为了分析温度对共享单车使用量的影响, 统计不同温度条件下每个小时内所有站点的单车借用量, 如图 5 所示. 可以发现: 当温度低于 47 华氏度时, 单车需求量很低, 最高不超过 700 辆; 当温度超过 47 华氏度(8.3 摄氏度)时, 用车需求明显增多, 单车使用量大部分维持在 1 000 辆以上; 但是随着温度的上升, 单车需求量有缓慢减少的趋势.

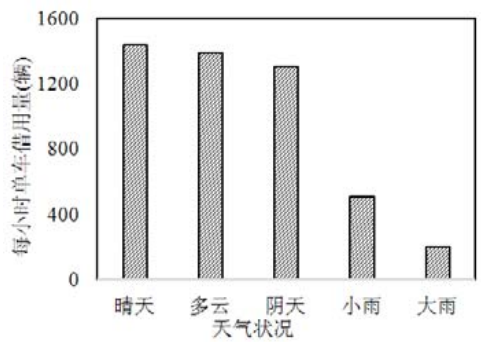


图 4 时间特征统计分析

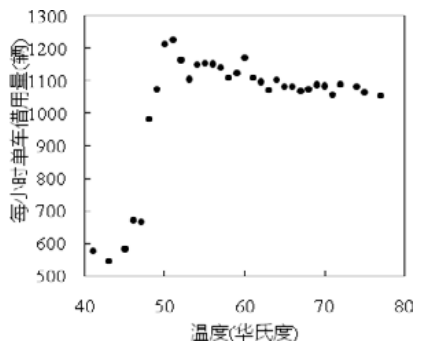


图 5 不同温度下每小时内单车借用量

(3) 湿度

同样, 湿度对共享单车的使用也会产生类似影响, 图 6 显示了不同湿度条件下每个小时内所有站点的单车借用量. 可以发现: 湿度在 34%–64%之间的用车需求最多, 平均借车量都在 1 000 辆以上; 而当湿度大于 64%时, 单车需求量明显减少, 单车需求量都在 1 000 辆以下.

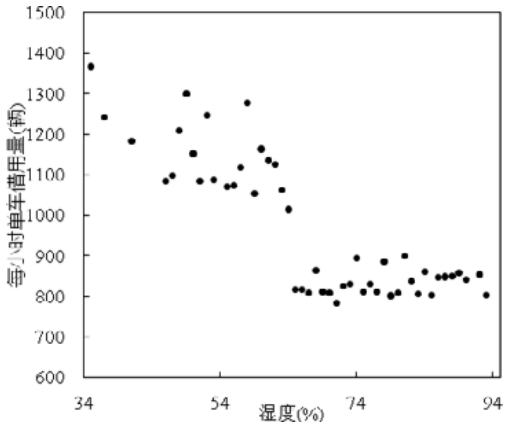


图 6　不同湿度下每小时内单车借用量

分析时间特征和天气特征对共享单车需求量的影响, 并对结果进行可视化分析, 得出的结论如下.

(1) 用户在工作日对共享单车的需求比双休日多, 且借用单车的数量波动不明显; 而用户在双休日对单车的使用量波动较大, 具有很强的随机性;

(2) 用户在工作日使用单车的高峰期集中在上下班的过程 8:00–10:00 和 18:00–20:00, 而在双休日对单车的需求平均分布在 9:00–19:00 之间;

(3) 不同的天气状况对用户使用单车的影响极大: 如果天气比较好, 单车的使用频率很高; 相反, 如果遇到下雨等极端天气, 对单车的需求量会急剧下降;

(4) 不同温度也会影响用户对单车的需求: 如果温度太低, 对单车使用的影响会尤为明显; 与之相比, 用户更愿意在暖和的条件下租用单车出行;

(5) 湿度同样会对单车的使用造成类似影响: 如果湿度不在人体的舒适范围内, 单车的需求量也会相应地有所降低.

除了时间和天气特征因素, 影响共享单车需求量的客观因素还包括:

1) 人口密度. 对人口密度进行量化, 首先需要确定人口密度函数和指定区域人口的数量和区域的有效面积, 在得到的人口数据归一化之后, 通过设计人口密度程度的计算公式, 获得某一区域的人口密度值;

2) 交通拥堵程度. 需要结合当地城市交通状况设计不同区域交通拥堵评价指标模型, 进而计算获得某一区域的交通拥堵情况. 由于城市交通数据的获取比较困难, 当前主要采用模拟的方法近似估计某一特定区域的交通拥堵程度;

3) 用户出行总成本. 用户出行总成本包括用户出行时间成本和用户出行货币成本. 对于用户出行时间成本, 可以结合用户到最近站点的距离并利用交通拥挤程度来进行估算. 在获得用户出行的时间成本和货币成本之后, 需要借助现有的出行方式平衡模型获得最终的出行总成本.

本文所采用的客观因素同样会影响到站点的分布, 如天气特征因素, 如温度、湿度、阵风速率、可见度、风速等因素对于不同区域站点设置的数量及密度会产生很大的影响. 例如: 雨水较多的区域, 单车站点的数量较少; 温湿度适宜的区域通常站点数量较多. 时间因素同样会对站点的分布产生影响, 通常在不同的月份, 用户对单车的需求不同, 直接导致站点数量产生较大变化. 因此, 在单车需求较旺盛的月份, 站点的数量会有所增加. 人口密度和交通拥堵程度情况同样会影响站点的数量. 通常, 在大城市人口密度较大、交通比较拥堵的区域, 站点的数量分布比较密集; 而人口稀疏或者交通状况较好的区域, 站点的分布比较稀疏.

3.3 特征矩阵表达

在 LSTM 神经网络模型中, 为了更好地获得准确的预测结果, 仅仅通过原数据无法立即呈现出其中包含

的规律, 所以模型中将第 3.1 节和第 3.2 节选取的动态特征转化为矩阵形式, 作为输入数据, 构建的二维特征矩阵, 见公式(4).

定义 7(特征矩阵).

$$X=\begin{pmatrix} X_1^1 & X_1^2 & X_1^j \\ \vdots & \ddots & \vdots \\ X_i^1 & X_i^2 & X_i^j \end{pmatrix} \tag{4}$$

其中, 特征矩阵的每个元素 X_i^j 表示第 i 个站点的第 j 个动态特征所对应的数值, i 表示单车行程记录中的站点编号, j 表示选取的特征编号. 将动态矩阵输入模型后, 预测结果为每个站点在不同因素下所对应的单车需求量. 下面以 3 个站点为例, 给出模型中输入的动态特征信息, 如下所示:

$$X=\begin{bmatrix} \text{站点编号} & \text{年份} & \text{月份} & \text{时刻} & \text{星期} & \text{温度} & \text{湿度} & \text{降风速率} & \text{可见度} & \text{风速} \\ 1 & 2014 & 9 & 15 & 1 & 66 & 73 & 5 & 10 & 8 \\ 3 & 2015 & 2 & 8 & 6 & 58 & 67 & 2 & 10 & 3 \\ 11 & 2015 & 3 & 14 & 4 & 61 & 66 & 5 & 10 & 4 \end{bmatrix}.$$

4 基于数据场的站点聚类方法

4.1 工作原理

借鉴 Google 网页排序算法 PageRank[10]的思想, 考虑站点间单车转移的数量, 得到所有站点的活跃度. 其次, 利用基于数据场聚类[11]的思路, 同时考虑站点位置和单车转移模式, 对共享单车系统内站点进行聚类. 其基本工作原理为: 1) 由历史行程记录和站点分布数据构建出单车转移网络, 根据站点间单车转移量, 分别考虑借车站点对自身活跃度影响和还车站点对借车站点活跃度的影响, 通过不断迭代计算站点活跃度; 2) 利用轮廓系数[12]评估方法, 确定前 q 个活跃度最高的站点作为簇中心, 根据单车转移网络, 构建在不同时间段每个站点的单车转移矩阵; 3) 考虑站点位置和单车转移模式, 对站点进行二级聚类, 进而得到地理位置接近且用车模式相似的站点聚簇. 基于数据场的站点聚类算法工作原理如图 7 所示, 具体内容见第 4.3 节算法 1.

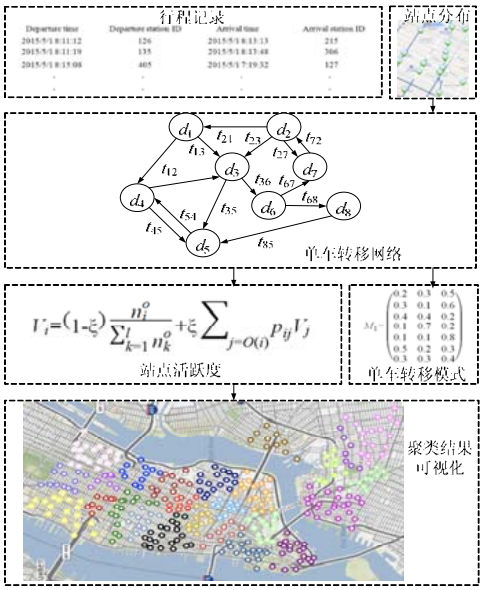


图 7 站点聚类算法原理图

借助站点聚类算法, 可以去除掉用户访问比较稀疏的站点信息, 对站点通过聚类后划分成不同的聚簇, 从而不会出现大量数据集中在某一到二个区域中的情况, 使站点的划分更加精细, 进而减少了潜在的数据倾

斜产生的影响.

4.2 站点活跃度

在共享单车系统中, 站点间的单车转移网络可以表示为一个加权有向图网络. 在该网络内, 每个顶点代表一个站点, 每条有向边表示两个站点间存在单车转移, 边上的权值表示两个站点间的单车转移数量. 借鉴 PageRank 算法的思想: 基于互联网里的网页被引用量计算该网页的重要性, 进而对网页排名. 因此, 单车站点活跃度可以类比为网页重要性, 如果很多站点内单车都是从站点 d_i 转移来的, 或者一个活跃度很高的站点 d_j 内有单车从站点 d_i 转移而来, 则站点 d_i 的活跃度 V_i 会因此增大.

在单车转移网络中, 从站点借出的单车数量越多, 则表明该站点的活跃度越高. 站点的活跃度主要受两个方面的影响: (1) 考虑借车站点对自身的活跃度影响, 借出的单车数量越多, 该借车站点的活跃度相应地也越高; (2) 考虑还车站点对借车站点的活跃度影响, 得到每个还车站点 d_j 分别对其借车站点 d_i 的活跃度贡献, 进而获得借车站点的活跃度.

本文采用幂迭代法得到收敛后的站点活跃度. 用 M_{out} 表示借车站点对自身活跃度产生的影响矩阵, M_{in} 表示还车站点对借车站点产生的活跃度矩阵, 得到如公式(5)所示的所有站点活跃度矩阵 H:

$$H=(1-\varphi)\times M_{out}+\varphi\times M_{in} \tag{5}$$

例 2(单车转移网络示例): 图 8 是一个含有 d_1~d_6 这 6 个站点的单车转移网络, 当有单车从站点 d_1 转移到站点 d_3, 则添加一条从 d_1 到 d_3 的有向边, 边上的数字表示转移单车的数量, 如: 有 5 辆单车从站点 d_1 转移到站点 d_3.

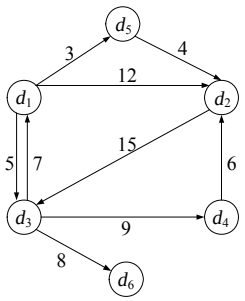


图 8 单车转移网络示例

图 8 中, 第 k 次迭代时各站点活跃度为 v_k, M 表示单车转移矩阵. 假设所有站点初始活跃度取为 0.25, 阻尼因子 φ 取值为 0.8, 通过公式(6)不断迭代, 当满足收敛条件后, 得到最终的站点活跃度:

$$v_{k+1}=M\times v_k \tag{6}$$

4.3 二级聚类算法

单车需求预测前需要对单车站点聚类, 主要原因在于:

(1) 每个站点的单车使用量受多种因素影响[14,15,20], 如天气、时间和站点间关联等, 使得单车使用量波动较大, 难以发现周期性和规律性, 因此无法保证预测的准确性. 经过聚类, 特征相似的站点聚集到一个聚簇, 其周期性和规律性更为明显, 相比于单个站点, 对单车需求的预测准确性也得以提高;

(2) 从现实因素考虑, 一个城市设立有很多站点, 且多数站点间距离很近, 如果用户无法在一个站点借/还车, 选择在附近站点借/还车也很方便;

(3) 如果遇到突发情况, 通常影响的是一个区域范围, 此时, 预测单个站点的单车需求并不能满足用户需求.

基于数据场的聚类算法思想来源于物理学势场的概念, 在复杂网络中, 各个节点间的影响是相互的, 节点在网络中重要性越大, 则其影响范围也越大. 因此, 可以把复杂网络类比为一个势场, 聚类时, 将节点加入到势最大的簇中心所在聚簇. 在单车转移网络中, 站点 d 对站点 d_i 所产生的势 $\chi_i(d)$ 的计算方法如公式(7)所示:

$$\chi_i(d)=v_i\times \mathrm{e}^{-\left(\frac{\|d-d_i\|}{\varepsilon}\right)^2} \tag{7}$$

其中, $\|d-d_i\|$表示站点 d 到站点 d_i 的距离; ε表示影响因子, 调节站点间的相互影响程度.

本文提出一种基于数据场的二级聚类算法 DFTLC (data field based two-level clustering algorithm), 考虑了站点位置信息, 方便用户在距离相近的站点借/还车. 此外, 算法考虑了单车转移模式[21−24], 使得对各个聚簇内单车的需求预测更为精确. 聚类过程如算法 1 所示, 主要步骤包括:

(1) 基于第 4.2 节介绍的内容, 利用第 4.4 节给出的轮廓系数评估方法确定前 q 个活跃度最高的站点作为簇中心, 基于站点地理位置, 按照势场大小, 将其他站点加入到这 q 个簇中心所在聚簇(第 1 行);

(2) 对每个站点生成单车转移矩阵(第 4–6 行);

(3) 同时考虑站点位置和单车转移矩阵, 对单车站点进行第 2 次数据场聚类(第 7 行);

(4) 重复步骤(2)、步骤(3), 如果聚类结果与上次相同或者达到迭代次数, 则聚类结束, 返回结果(第 8 行−第 11 行).

算法 1. 基于数据场的二级聚类算法.

输入: 单车站点 $\{d_i\}_{i=1}^n$, 历史行程记录 $\{Trj_i\}_{i=1}^T$, 总迭代次数 K, 迭代次数初始值 k=0;

输出: q 个站点聚簇: $\{C_{K,1},C_{K,2},\ldots,C_{K,q}\}$.

1. 基于站点位置对 $\{d_i\}_{i=1}^n$ 利用数据场聚类算法得到 q 个聚簇$\{C_{k,1},C_{k,2},\ldots,C_{k,q}\}$;
2. **while** $k<K$
3. $k=k+1$;
4. **for** i=1 **to** n
5. 对站点 d_i 生成单车转移矩阵 M_i;
6. **end for**
7. 基于站点地理位置信息对所有站点 $\{d_i\}_{i=1}^n$ 应用数据场聚类算法, 得到$\{C_{k,1},C_{k,2},\ldots,C_{k,q}\}$;
8. **if** $\{C_{k,1},C_{k,2},\ldots,C_{k,q}\}=\{C_{k-1,1},C_{k-1,2},\ldots,C_{k-1,q}\}$
9. **return**;
10. **end if**
11. **return** $\{C_{K,1},C_{K,2},\ldots,C_{K,q}\}$;

- 算法时间复杂度分析.

DFTLC 算法主要包括两个阶段.

- 第 1 阶段是为每个站点找距其最近的簇中心(具体参加第 4.4 节内容), 并将其划分给该簇中心所对应的聚簇, 时间复杂度为 O($n\times q\times k$), 其中, k 为迭代次数, 初始值设置为 10 次, 因为算法运行过程中, 迭代次数一般不大于 10 次便停止迭代;
- 第 2 阶段是对每个站点生成单车转移矩阵, 同时考虑两个因素进行二级聚类, 时间复杂度为 O($n\times q\times k$). DFTLC 算法的时间复杂度为 O($n\times q\times k$).

例 3(基于数据场的二级聚类示例): 图 9 描述了 8 个单车站点采用基于数据场的二级聚类迭代过程, d_1, d_4, d_6 为 3 个初始簇中心站点.

在 DFTLC 算法的一级聚类中只考虑站点位置, 将剩余的站点划分到距离最近的簇中心所在聚簇, 得到 $\{C_{0,1},C_{0,2},C_{0,3}\}$; 在二级聚类中同时考虑站点位置和单车使用模式, 由于簇中心 d_4 对站点 d_5 产生的势更大, 故将 d_5 重新划分到 d_4 所在聚簇. 二级聚类过程迭代 K 次, 直至达到迭代终止条件.

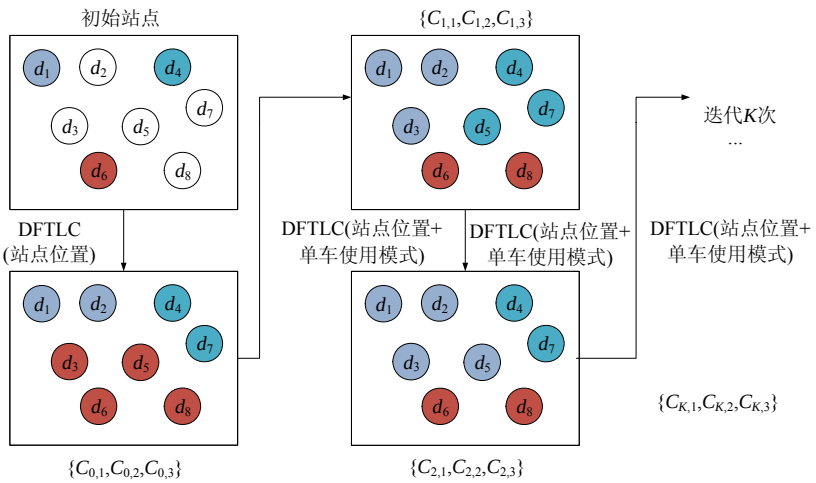


图 9 基于数据场的二级聚类分析

4.4 簇中心选取

在第 4.3 节中, 簇中心站点的个数 q 是预设的经验值, 不同的 q 值会产生不同的聚簇. 为了确定最佳的簇中心个数 q 值进而得到最佳的聚类结果, 选取轮廓系数作为聚类效果好坏的依据, 使聚簇内站点的相似度尽可能大, 聚簇间的站点相似度小.

假设站点 i 被划分到聚簇 C_i, 聚簇内站点不相似度由站点 i 到聚簇 C_i 内其他站点的平均距离 $a(i)$决定, $a(i)$越小, 则站点聚类效果越好. 该聚簇 C_i 内所有站点的 $a(i)$均值为该聚簇 C_i 的不相似度.

聚簇间不相似度为站点 i 到其他聚簇 C_j 到的所有站点的平均距离 b_i(即站点 i 与聚簇 C_j 的聚簇间不相似度). 站点 i 的聚簇间不相似度如公式(8)所示:

$$b_i=\min\{b_{i,1},b_{i,2},\ldots,b_{i,q}\}\ (q\neq i) \tag{8}$$

$b_{i,q}$ 表示站点 i 与聚簇 C_q 的不相似度, b_i 越大, 表明站点 i 越不应该属于其他聚簇.

轮廓系数可表示为公式(9)、公式(10):

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{9}$$

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases} \tag{10}$$

由公式(10)可以发现: $s(i)$的范围在(–1,1)之间, 越接近于 1, 表示站点 i 的被划分到正确的聚簇; 相反, 如果取值越接近于–1, 表示站点 i 的被划分到错误的聚簇; 如果 $s(i)$的值为 0, 表示站点 i 被划分到两个聚簇的边界上.

因为预测系统轮廓系数是所有站点轮廓系数 $s(i)$的平均值, 对于聚类效果的评价具有一定的有效性. 当轮廓系数最大时的 q 值为最佳的簇中心个数.

5 共享单车需求预测模型

考虑需求量预测问题时, 需要确定与单车需求相关的数据特性, 进而分析这些数据对预测结果的影响. 单车需求量的变化与时间和天气有关, 而且天气数据包含许多特征, 因此首先分析时间和天气数据对共享单车需求的影响, 然后将它们作为关键因素, 预测在不同时刻不同天气条件下的单车需求量.

5.1 工作原理

本文利用长短时记忆网络 LSTM 构建了一个多因素 LSTM 网络模型, 提出了 DeepML (deep multi-features LSTM network)算法预测单车需求, 以 30 分钟为长时间间隔, 对每个聚簇内的单车需求进行预测. 基于站点聚类的单车需求预测算法原理包含 4 个部分, 如图 10 所示.

(1) 数据采集层: 用户每次使用共享单车后, 系统将保存用户的行程记录信息, 进而分析每个聚簇每半小时内对单车的需求;

(2) 数据预处理层: 分析天气和时间数据对单车需求的影响, 将第 4.2 节分析得到的对单车使用量有影响的多个特征表达为向量形式;

(3) 多因素 LSTM 网络层: 将特征向量输入到多因素 LSTM 网络模型中, 使用共享单车需求量的时间序列对模型进行训练, 得到模型最优参数, 预测每个聚簇半小时后所有站点对单车的需求;

(4) 数据可视化层: 对单车需求预测结果可视化输出并与原始数据对比分析.

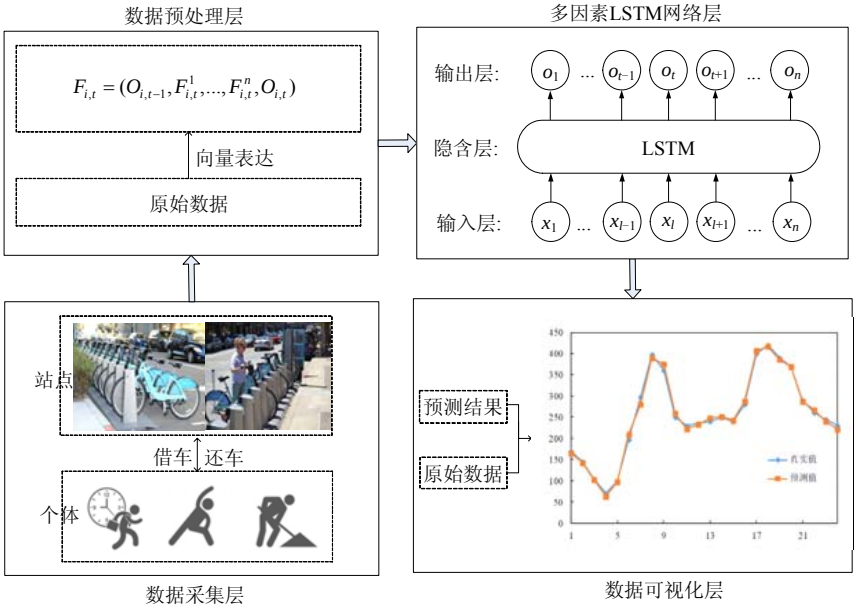


图 10 单车需求预测模型工作原理

5.2 DeepML算法

多特征 LSTM 网络与传统神经网络模型一样, 使用的是多层网络结构. 由于多特征 LSTM 网络引入了记忆细胞的概念, 信息在记忆细胞的水平线上传输和处理, 可以长时间存储信息, 因此非常适用于单车需求预测问题. 单车需求预测算法 DeepML 利用多特征 LSTM 网络预测单车使用量, 包括一个输入层、两个 LSTM 层、一个全连接层和一个输出层, 利用激活函数和 dropout 两个关键操作, 通过 dropout 技术控制模型的规模, 预测模型的泛化能力也有所增强.

(1) 输入层

在时间序列预测模型中, 为了获得准确的预测结果, 仅仅通过原数据无法呈现出其中包含的规律, 所以模型中将第 4.2 节选取的特征归一化再转化为向量形式, 作为输入数据. 以当前时间段 t 内天气特征和前一个时间段$\{t-1\}$内借车数量来预测下一个时间段$\{t+1\}$内借车的数量为例, 构建的特征向量, 如公式(11)所示:

$$F_{i,t}=(O_{i,t-1},F_{i,t}^1,...,F_{i,t}^n,O_{i,t}) \tag{11}$$

其中, $F_{i,t}$ 表示在当前时间段 t 内聚簇 C_i 的特征向量, 包括上一时间段$\{t-1\}$内聚簇 C_i 的借车数量 $O_{i,t-1}$、当前时

间段 t 内聚簇 C_i 的 n 个天气特征 $F_{i,t}^n$ 和当前时间段 t 内聚簇 C_i 的借车数量 $O_{i,t}$.

在单车需求预测问题中的输入是三维向量(*samples*, *timesteps*, *features*), 其中, *samples*, *timesteps*, *features* 分别表示训练样本数量、时间步长和 $F_{i,t}$ 的特征向量. 在多特征 LSTM 网络模型中输入三维向量, 使用最近 l 个时间段状态信息, 即 *timesteps*=l, 预测在不同天气状况下每个聚簇下一时间段内的单车需求. 本文将单车需求预测问题形式化表达为下式:

$$O_{i,t+1}=DeepML(F_{i,t-l},\ldots,F_{i,t-1},F_{i,t}) \tag{12}$$

本文提出的多因素 LSTM 网络模型的输入融合了高维特征, 即第 3 节讨论的多维度的天气和时间特征向量, 利用特征向量 $F_{i,t}$ 表达, 具体包括天气因素(温度、湿度、阵风速率、可见度、风速)、时间因素(年份、月份、时刻、星期), 这样可以保证有效融合影响用户借还车的客观因素, 进而更加准确地预测用户借还单车的数量.

如果有新的特征考虑, 如人口密度、交通拥堵程度、用户出行总成本, 可以将这些新特征作为不同维度信息, 加入到公式(11)定义的 $F_{i,t}$ 特征向量.

(2) LSTM 层

在未来一段时间, 用户对聚簇内单车的需求量会受当前以及前一段时间单车站点状态的影响. 因此, 为了记住较远时间以前站点的状态, 本文分别研究了前 4 个时间步对下一时刻聚簇内单车需求量的影响, 其中每个时间步大小设为 30 分钟. 此外, 设置了两个 LSTM 层, 每层包含 50 个神经元.

对于多特征 LSTM 网络的构建, 本文使用 keras 框架中的 *Sequential*(·)方法对每个 LSTM 层进行堆叠, 除了最后一层 LSTM 的参数 *return_sequence* 需要设置为 False, 前 N–1 层的 *return_sequence* 都要设置为 True, 进而保证向下一层传播在当前时间步长上的单车需求预测值.

由于构造的时间序列通常都会存在自相关性问题, 即预测模型往往把{t–1}时刻的单车需求量作为 t 时刻单车需求量的预测值, 即使最后的预测误差看起来很小, 实际上真实误差依旧很大. 在对单车数据进行分析时, 发现单车需求量以 7 天为一个周期波动. 为了消除序列的自相关性, DeepML 算法根据序列的周期性进行差分运算, 得到平稳的时间序列后, 将当前时刻与上一时刻的单车需求量差值作为回归目标.

为了使预测模型能够学习到训练样本之间的时序特征, 多特征 LSTM 网络采用有状态的记忆网络, 在对样本进行训练的过程中, 保存样本批(batch)之间的状态信息, 将当前批的训练样本状态值作为下一批训练样本的初始隐藏状态, 使得每个样本批之间存在信息传递. 构建保存样本批之间状态的 LSTM 网络需要保证使用的训练样本能反映数据的周期性, 比如单车需求变化的周期是 7 天, 并在每次训练完一个样本批之后需要重置 LSTM 状态.

在多特征 LSTM 网络中, 使用 *sigmoid* 函数构造多个控制门实现对信息的筛选. 使用 *softsign* 函数代替了传统的 tanh 函数. *softsign* 函数与 tanh 函数特性很相似, 但由于更容易饱和, 所以训练速度更快.

在消除了时间序列之间存在的自相关性基础上, 通过构建保存样本批状态信息的多特征 LSTM 网络结构, 前 l 个时间步的单车需求量将作用于当前时间段, 进而实现序列的长时间记忆. 多特征 LSTM 网络结构如图 11 所示.

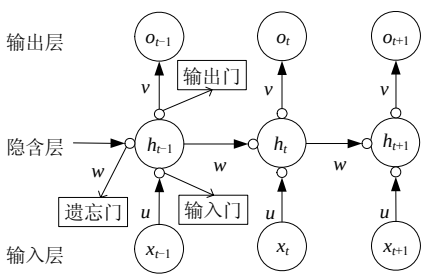


图 11 多特征 LSTM 网络结构

当前时间段隐藏状态 h_t 不仅与当前时间段的输入 x_t 有关, 而且与前 l 个时间段的状态有关. h_t 可表示为

$$f(u\times x_t+w_{t-1}\times h_{t-1}+w_{t-2}\times h_{t-2}+\ldots+w_{t-l}\times h_{t-l}).$$

其中, w 和 u 表示权重矩阵, h_{t-l} 表示隐藏层的中间状态. 进而实现与前面 l 个时间段的状态关联.

(3) 全连接层

在计算单车需求量的真实值与预测值之间的损失值前, 需要先经过全连接层对 LSTM 层的输出进行维度变换, 使得输出向量的维度与真实值的向量保持一致. 全连接层将当前层的结点与上一层结点相连, 综合前面提取的特征, 进行共享单车需求预测.

多因素 LSTM 网络包含一个全连接层, 对 LSTM 层得到的结果进行降维处理后, 只保留有用信息. 因为本文研究的是单车需求量预测值, 所以经过全连接层处理后, 输出的维度为 1. 此外, 通过 dropout 技术从输出结果中减少一定比例的元素对预测效果的影响, 防止过拟合现象.

(4) Dropout 技术

本文将 dropout 概率值 p 设为 0.3, p 为大量实验获得的经验值. 通过实验发现: 当 p=0.3 时, 可以使损失函数值最小. 在训练时, 每个神经元按照 0.3 的概率被隐藏, 这部分神经元在前后传播中均保持不变, 暂时不更新权重. 在测试时, 每个神经元总是处于激活状态, 每个神经元的权重均乘以 0.7(即 1−0.3), 以确保神经元的输入期望值与训练过程中该神经元的期望相同.

(5) 输出层

因为多因素 LSTM 网络的训练样本经过归一化处理, 所以输出的结果也是经过归一化的标准值, 对输出结果逆转换后的值, 即为每个聚簇在不同时间段、不同天气状况下的单车需求量. 全连接层的输出作为输出层的输入, 由全连接层的输出与输出层权重和偏置项计算得到最终预测结果.

5.3 算法复杂度分析

DeepML 算法的参数类型有 4 种, 分别是遗忘门、输入门、输出门和细胞状态, 故参数总数为 $4\times((d+r)\times r+r)$, 其中, d 表示输入特征的维数, r 表示隐藏层神经元的个数. 算法的主要操作是对参数进行更新, LSTM 使用反向传播的方法更新参数, 每次参数更新的时间复杂度为 $O(a\times b+a\times c+b\times s+c\times s)$, 其中, a 表示输出单元的个数, b 表示隐藏单元的个数, c 表示记忆元件的大小, s 表示记忆元件、隐藏单元和门控单元直接相连的单元的个数.

5.4 优化算法

优化算法的作用是改善模型的训练方式, 从而最小化或者最大化目标函数. 梯度下降是在神经网络中最常用的优化算法, 常用的算法包括随机梯度下降(SGD)、RMSProp[25]、Ftrl[26]、Adagrad[27]等. 本文使用的是梯度下降算法的一种变体 Adam (adaptive moment estimation)[28], 该算法结合了 Momentum 和 RMSprop 算法的思想, 可以计算每个参数的自适应学习率, 为每个参数动态调整学习率. 在实际应用中, 该算法效果较好, 相比于其他自适应学习率算法, 收敛速度更快. Adam 算法优化过程如下.

公式(13)用于计算第 t 次梯度的一阶矩(期望)m_t, 其中, g_t 是损失函数对未知参数 θ 第 t 次求导得到的梯度, β_1 表示一阶矩的衰减系数:

$$m_t=\beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t \tag{13}$$

公式(14)用于计算第 t 次梯度的二阶矩 v_t, 其中, g_t^2 表示 g_t 的 Hadamard 积, β_2 表示二阶矩的衰减系数:

$$v_t=\beta_1\cdot v_{t-1}+(1-\beta_1)g_t^2 \tag{14}$$

公式(15)用于计算 m_t 的偏置矫正 m_t', 其中, β_1^t 表示第 t 次一阶矩的衰减系数 β_1.

$$m_t'=\frac{m_t}{1-\beta_1^t} \tag{15}$$

公式(16)用于计算 v_t 的偏置矫正 v_t', 其中, β_2^t 表示第 t 次二阶矩的衰减系数 β_2.

$$v_t' = \frac{v_t}{1-\beta_2^t} \tag{16}$$

公式(17)表示对未知参数θ更新的过程:

$$\theta_t = \theta_t - 1 - \alpha * \frac{m_t'}{(\sqrt{v_t'} + \varepsilon)} \tag{17}$$

其中,

- ε值表示步长, 通常设置为极小值, 防止出现分母为 0 的情况;
- α是学习率, 控制参数每次更新的速度: 如果学习率设置太大, 由于学习速度过快, 可能会导致参数在最优值两侧来回震荡, 甚至无法在有限的时间内收敛到局部最优值; 如果设置太小, 收敛过程将变得十分缓慢.

由上述公式得知, 需要确定一阶矩的衰减系数β_1、二阶矩的衰减系数β_2、学习率α和ε值. 在每次的迭代训练过程中, 学习率都在一个确定的范围内, 参数的变化比较平稳, 当满足终止条件时, 就能得到最优参数的解. 在本文中, Adam 算法 3 个参数β_1, β_2, α和ε分别初始化 0.9, 0.999, 0.001, 10^{-8}.

6 实验分析

6.1 对比算法

实验中, 将本文提出的基于数据场的二级聚类算法 DFTLC 分别与 *k*-means 算法和不考虑单车转移模式的基于数据场的二级聚类算法 DFTLC-α 算法进行比较. 为了验证本文所提预测模型的准确性, 将本文提出的 DeepML 预测算法与以下 3 种算法进行对比实验.

(1) HA (historical average): 使用历史对应时间段内单车借/还量的平均值作为预测结果, 例如预测周一 7:00–8:00 单车借出数量, 则用历史的工作日对应时间段内单车借出数量的平均值作为预测值;

(2) GBDT (gradient boosting decision tree)[29]: 通过多次迭代生成多个单车需求回归树, 每个学习器的训练在上一轮学习器的残差基础上进行, 将得到的多个学习器进行加权求和, 得到最终的预测结果;

(3) HP-*K*NN (hierarchical prediction-*K* nearest neighbor): 该算法首先用于预测单车全部借/还数量; 然后预测聚簇间用车比例, 将全部用车量按此比例分配到每个聚簇; 最后, 基于 *K*NN 算法[30]作出层次预测.

6.2 实验数据及环境

6.2.1 共享单车数据

实验中使用纽约 Citi Bike 系统 2014 年 4 月 1 日–9 月 30 日共享单车行程记录数据(https://www.citibikenyc.com/system-data/). 由 6 719 辆共享单车在 337 个站点间形成的 5 359 995 条单车转移记录, 数据格式为(行程时间、出发时间、到达时间、出发站点 ID、出发站点经/纬度、到达站点 ID、到达站点经/纬度). 本文选取 2014 年 4 月 1 日–9 月 10 日的行程数据作为训练集, 2014 年 9 月 11 日–30 日的行程数据作为测试集.

6.2.2 天气数据

实验中使用纽约 2014 年 4 月 1 日–9 月 30 日的天气数据, 来源于 mesowest (https://mesowest.utah. edu/), 记录了站点所在服务区域内的天气状况信息. 其中, 缺失的天气指标数值利用前一时刻的数据进行填充, 然后对其进行归一化处理. 原始共享单车数据并没有提供天气信息, 需要结合本节的天气数据一同使用, 预测单车数量.

6.2.3 实验环境

本文所有算法利用面向对象 Python 语言编程实现, 硬件平台为 2.8 GHz Intel Core i7 的 CPU, 内存为 16 GB, 运行在 Apple 的 OS X 操作系统上.

6.3 评价指标

为了评价预测性能的准确性, 本文使用均方根对数误差(root mean squared logarithmic error, RMSLE)和错误率(error rate, ER)作为评价指标. 上述指标用于衡量每个聚簇内用车数量的预测值与真实值之间的偏差, 定义如公式(18)、公式(19)所示:

$$ER=\frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{i=1}^{q}|O'_{i,t}-O_{i,t}|}{\sum_{i=1}^{q}O_{i,t}} \tag{18}$$

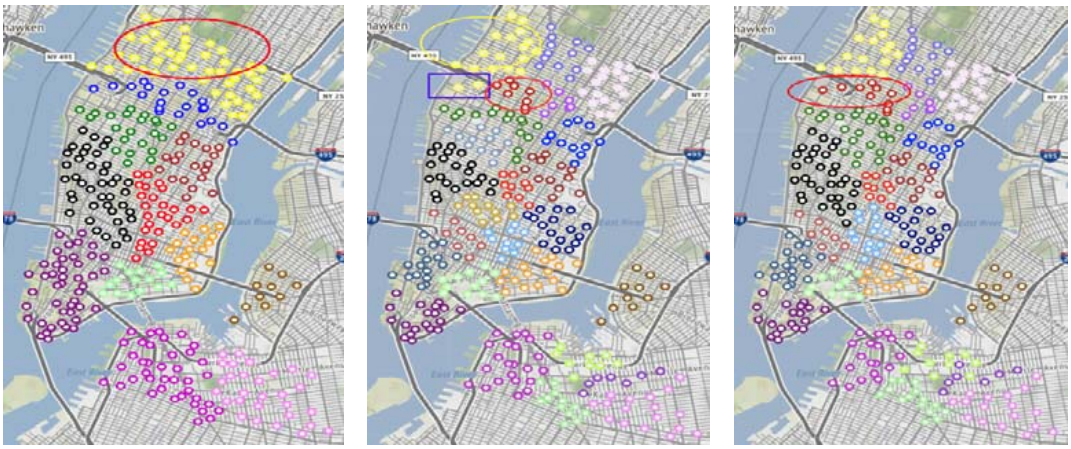
$$RMSLE=\frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{q}\sum_{i=1}^{q}(\log(O'_{i,t}+1)-\log(O_{i,t}+1))^2} \tag{19}$$

其中, T 表示时间戳的个数, q 表示聚簇个数, $O_{i,t}$ 表示 t 时间段内在聚簇 C_i 用户借车数量真实值, $O'_{i,t}$ 表示 t 时间段内在聚簇 C_i 内用户借车数量的预测值.

6.4 聚类算法性能分析

6.4.1 站点聚类结果对比

本实验使用基于数据场的二级聚类算法对站点进行聚类, 通过第 4.4 节介绍的轮廓系数指标评估法, 得到最佳聚簇个数分别为 12, 25 和 23 时, *k*-means, DFTLC-α和 DFTLC 这 3 种方法的聚类结果, 如图 12 所示.



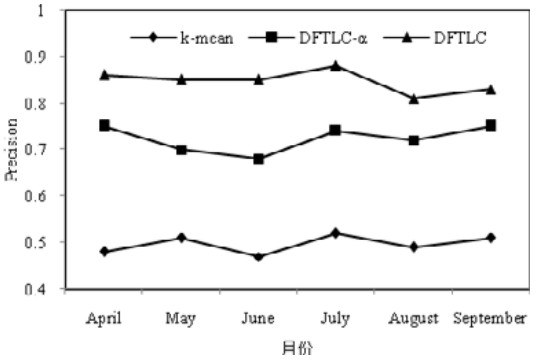
(a) *k*-means (b) DFTLC-α (c) DFTLC

图 12 站点聚类结果

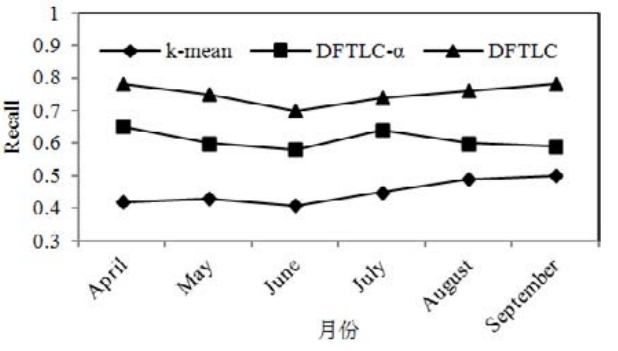
实验结果表明: 图 12(a)用 *k*-means 聚类得到的聚簇中站点数规模较大, 如图中椭圆圈内区域. 由于不同聚簇间距离较远, 一旦聚簇内单车数和停车桩数量不平衡, 在聚簇间进行单车调度将无法为用户提供便利. 图 12(b)给出了 DFTLC-α算法的聚类结果, 算法没有考虑单车转移模式, 以活跃度最高的 25 个站点作为簇中心, 因此聚簇内站点数相较于图 12(a)更少. 可以发现, 图 12(a)椭圆圈区域被划分为 3 个聚簇. 图 12(b)中有些单车转移模式相似的站点不在同一个聚簇, 如图中长方形区域两个站点的单车转移模式与右边椭圆圈标注的聚簇更相似, 然而被划分到上面的大的椭圆圈聚簇内. 图 12(c)则不会出现这一问题, 因为 DFTLC 算法同时考虑了单车转移模式和站点位置. 可以发现, 图 12(b)中长方形区域这两个站点被正确划分.

为了进一步证明所提基于数据场的二级站点聚类算法的性能优势, 本文从准确率 Precision、查全率 Recall 以及 *F*1-measure, 观察 2014 年 4 月−2014 年 9 月, 6 个月间 3 种算法的聚类结果, 如图 13 所示. 实验结果表明: 在 3 个聚类评价指标上, DFTLC 算法均高于其他两个算法: DFTLC 算法在准确率 Precision 上分别高于 *k*-means 和 DFTLC-α算法 35.0%和 12.3%, 在查全率 Recall 上分别高于 *k*-means 和 DFTLC-α算法 30.2%和 14.2%, 在 *F*1-measure 上分别高于 *k*-means 和 DFTLC-α算法 32.4%和 13.4%. 主要原因在于: (1) DFTLC 算法基于数据场

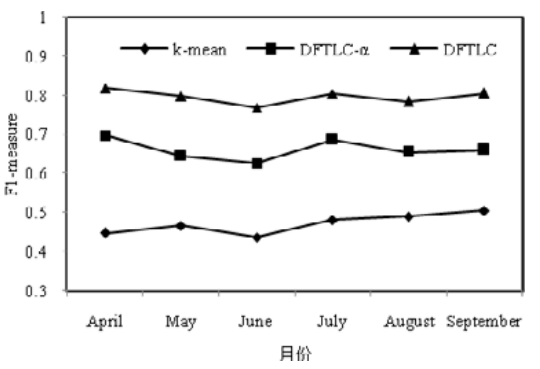
聚类的思想, 同时考虑了单车转移模式和站点的位置信息; (2) 相比于 *k*-means 算法初始聚簇个数的随机性, DFTLC-α和 DFTLC 选择轮廓系数最大时的 q 值作为最佳的簇中心个数, 使聚类结果更加接近于真实值, 聚类的准确性明显高于 *k*-means 算法.



(a) 聚类准确率对比



(b) 聚类查全率对比

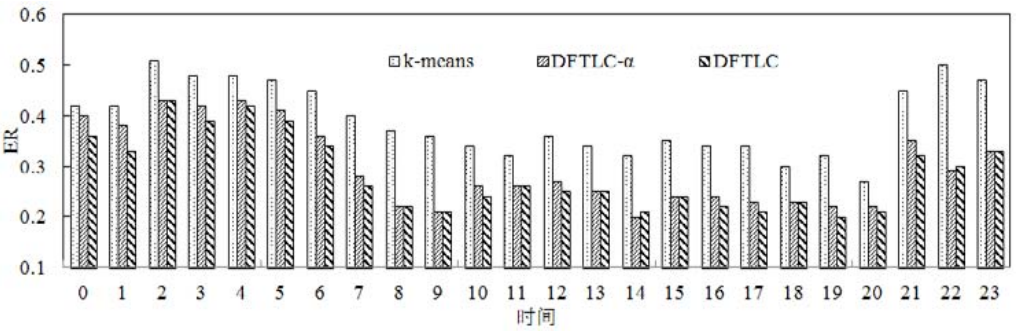


(c) 聚类 *F*1-measure 值对比

图 13 站点聚类准确性评价指标对比

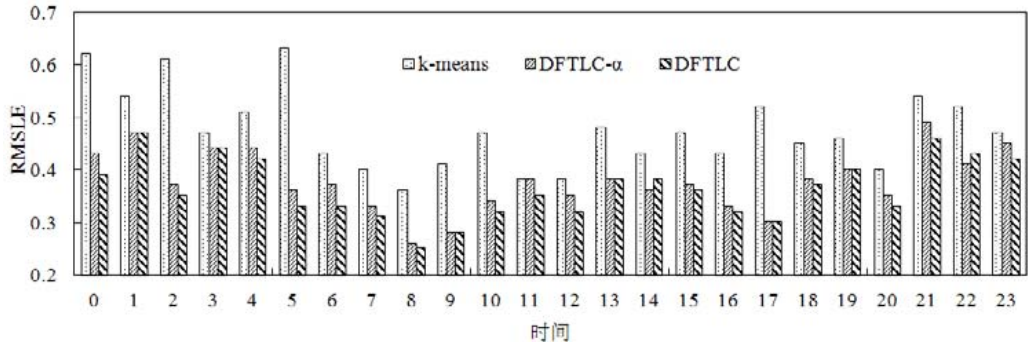
6.4.2 借车数量误差分析

图 14 给出了使用 3 种不同聚类算法在不同时间(小时)下, DeepML 算法预测借车数量的误差对比结果.



(a) 不同时间下 ER 值

图 14 不同算法借车数量误差对比



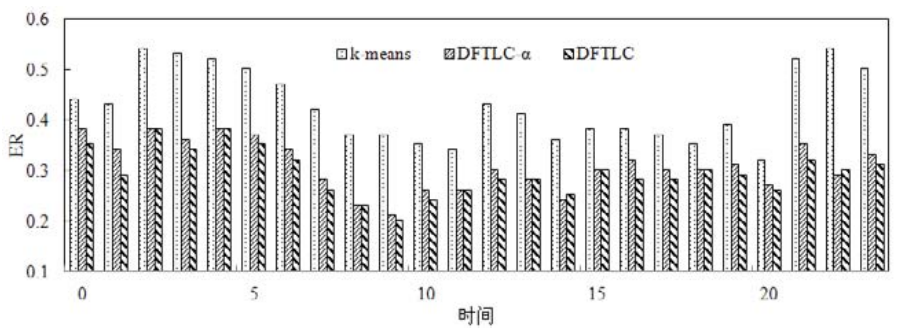
(b) 不同时间下 RMSLE 值

图 14 不同算法借车数量误差对比(续)

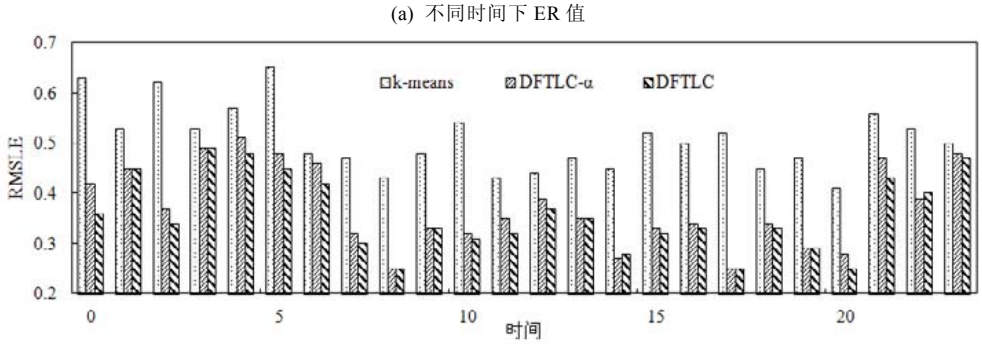
实验结果表明, 使用 *k*-means 聚类算法的预测误差最大. 通过计算各个站点的活跃度, 以活跃度高的站点作为簇中心, 基于站点位置进行聚类的 DFTLC-α算法, 相较于 *k*-means 算法预测效果有很大提升. 在此基础上, 同时考虑单车使用模式的 DFTLC 算法预测性能又进一步提升. 实验中发现, 在 21:00–5:00 期间的预测误差相对较大. 主要原因在于: 这个时间段内, 用户处于休息状态, 系统产生的单车行程记录数据较少. 相应地, 在白天活动期间, 由于 8:00–9:00 和 17:00–18:00 时间段处于上下班的高峰期, 产生大量行程记录数据, 因此预测误差也最小.

6.4.3 还车数量误差分析

图 15 表示了使用 3 种不同聚类算法在不同时间下, DeepML 算法预测还车数量的误差对比结果. 可以发现: 同时考虑站点位置和单车使用模式的 DFTLC 算法预测性能最佳, 原因同上一节类似.



(a) 不同时间下 ER 值



(b) 不同时间下 RMSLE 值

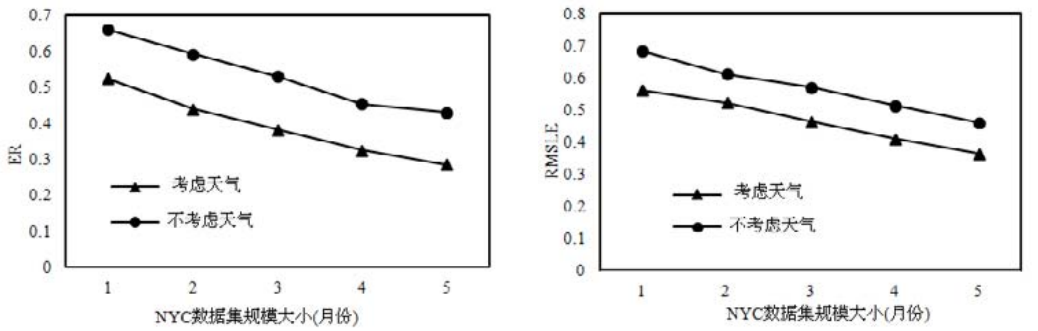
图 15 不同算法还车数量误差对比

6.5 预测算法性能分析

6.5.1 天气因素对预测性能的影响

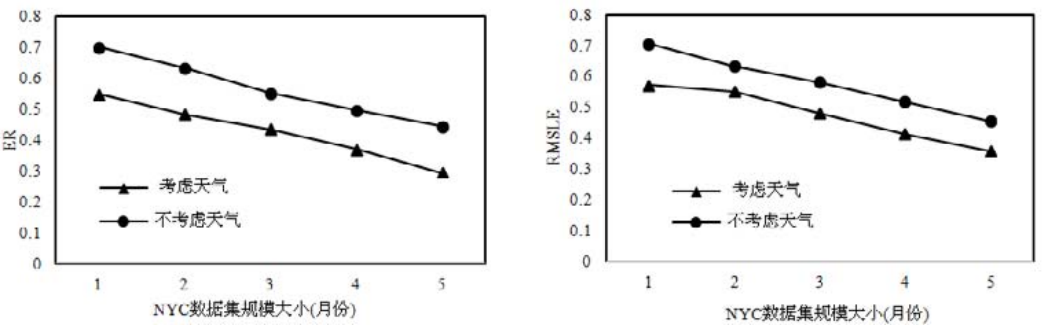
根据第 3.2 节对天气特征的分析可以发现, 天气因素对单车使用量具有显著影响. 图 16 和图 17 表示在不同数据规模下(按月份进行划分), 考虑天气因素和不考虑天气因素时借/还车数量预测性能的对比结果.

实验结果表明: 在不同数据规模下, DeepML 算法在考虑天气因素时, 相较于不考虑天气因素时具有更高的预测准确度. 以图 17 为例: 对于 ER 评估指标, DeepML 算法考虑天气因素比不考虑天气因素平均提高了 14.2%; 对于 RMSLE 评估指标, DeepML 算法考虑天气因素比不考虑天气因素平均提高了 10.4%. 原因如第 4.2 节所述, 因为天气会影响到用户的安全出行. 例如: 在晴天, 单车使用量比下雨天明显增多, 雨天用车条件恶劣, 比如道路积水易滑等.



(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 16 天气因素对借车数量预测性能的影响



(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 17 天气因素对还车数量预测性能的影响

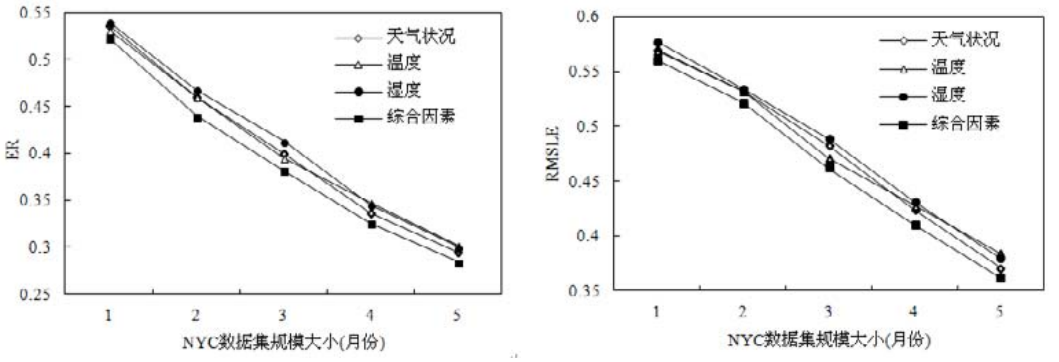
6.5.2 具体天气特征对预测性能的影响

根据第 6.5.1 节实验结果可以发现, 考虑天气特征能提高单车需求量的预测准确度. 本节分析具体天气特征对单车预测准确性的影响.

图 18 表示不同数据规模下(按月份进行划分)具体天气特征对借车数量预测的影响, 对于 ER 评估指标, DeepML 算法考虑综合天气因素比仅考虑天气状况因素平均提高了 3.7%, 比仅考虑温度因素平均提高了 4.1%, 比仅考虑湿度因素平均提高了 5.7%. 对于 RMSLE 评估指标, DeepML 算法考虑综合天气因素比仅考虑天气状况因素平均提高了 2.8%, 比仅考虑温度因素平均提高了 3.0%, 比仅考虑湿度因素平均提高了 4.1%.

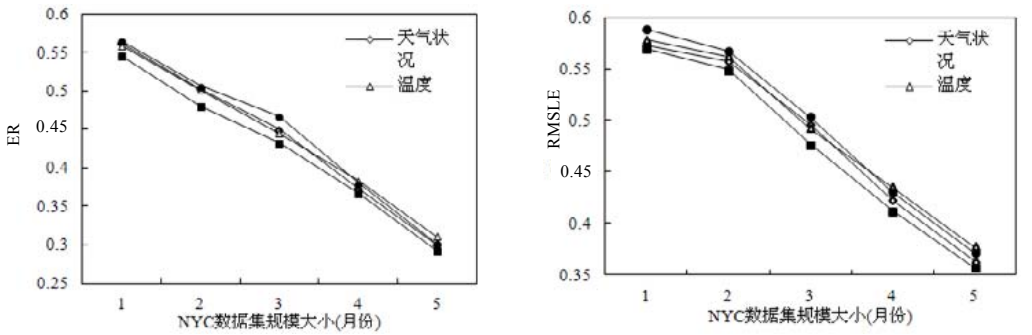
图 19 表示不同数据规模下(按月份进行划分), 具体天气特征对还车数量预测的影响. 对于 ER 评估指标, DeepML 算法考虑综合天气因素比仅考虑天气状况因素平均提高了 3.3%, 比仅考虑温度因素平均提高了 3.8%, 比仅考虑湿度因素平均提高了 4.8%. 对于 RMSLE 评估指标, DeepML 算法考虑综合天气因素比仅考虑天气状况因素平均提高了 2.0%, 比仅考虑温度因素平均提高了 3.3%, 比仅考虑湿度因素平均提高了 4.0%.

实验结果表明: 在不同数据规模下, DeepML 算法在考虑综合天气因素时, 相较于仅考虑单个天气因素时具有更高的预测准确度. 此外, 仅考虑天气状况因素比仅考虑温度、湿度因素的预测准确度都高. 因此, 天气状况对单车使用量影响最大.



(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 18 不同天气特征对借车数量预测性能的影响



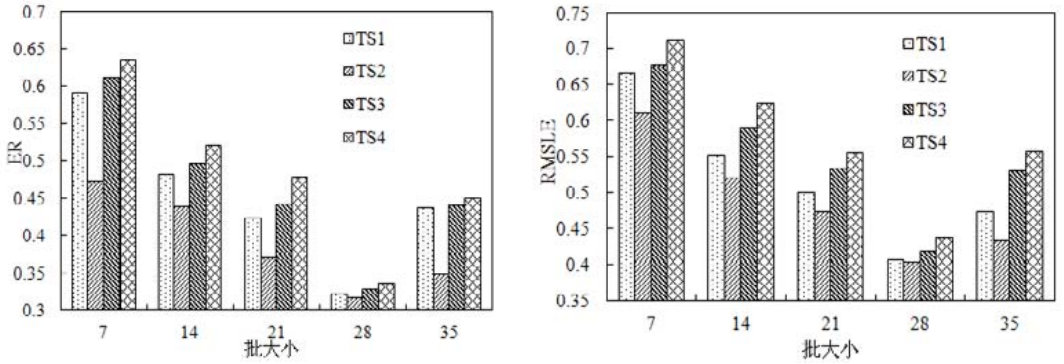
(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 19 不同天气特征对还车数量预测性能的影响

6.5.3 时间步对预测性能的影响

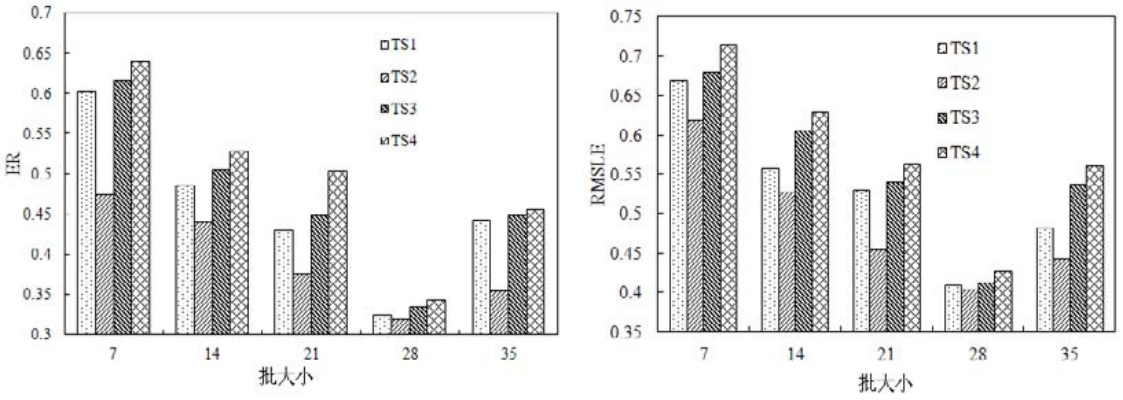
在共享单车系统中, 通常站点在某一时刻的状态受前 l 个时间段状态的影响, 为了实现序列的长时间记忆, 以 7 天为一个周期对训练样本进行批处理, 在多特征 LSTM 网络中保存样本批之间的状态信息, 使用反映单车借还周期性的批大小(即 7 的倍数)的数据来训练网络. 图 20 和图 21 表示在不同样本批大小下, 使用不同时间步时, 借/还车数量预测性能的对比结果. *TS*1 表示当前时间的前一个时间步, 依此类推.

实验结果表明: 当使用批大小为 28 进行训练时, DeepML 算法用前两个时间步预测单车需求量的准确度最高. 以图 20 为例: 对于 ER 评估指标, DeepML 算法用前两个时间步预测相较于前 1 个时间步平均提高了 25.6%, 相较于前 3 个时间步平均提高了 26.4%, 相较于前 4 个时间步平均提高了 29.0%; 对于 RMSLE 评估指标, DeepML 算法用前两个时间步预测相较于前一个时间步平均提高了 9.0%, 相较于前 3 个时间步平均提高了 22.1%, 相较于前 4 个时间步平均提高了 28.3%. 通过本节实验, 选取时间步为 2. 通常情况下, 考虑前 l 个时间段的站点状态对当前时间段内站点状态影响的步长越大, 对预测的准确性会越高; 但是当 l 达到某个一定的值时, 由于增加了对时序预测无关的噪声特征, 预测性能不升反降, 相反还增加了训练时间, 因此需要通过实验选取最佳时间步长.



(a) 不同时间步下 ER 值　　(b) 不同时间步下 RMSLE 值

图 20　时间步对借车数量预测性能的影响



(a) 不同时间步下 ER 值　　(b) 不同时间步下 RMSLE 值

图 21　时间步对还车数量预测性能的影响

6.5.4　优化算法对预测性能的影响

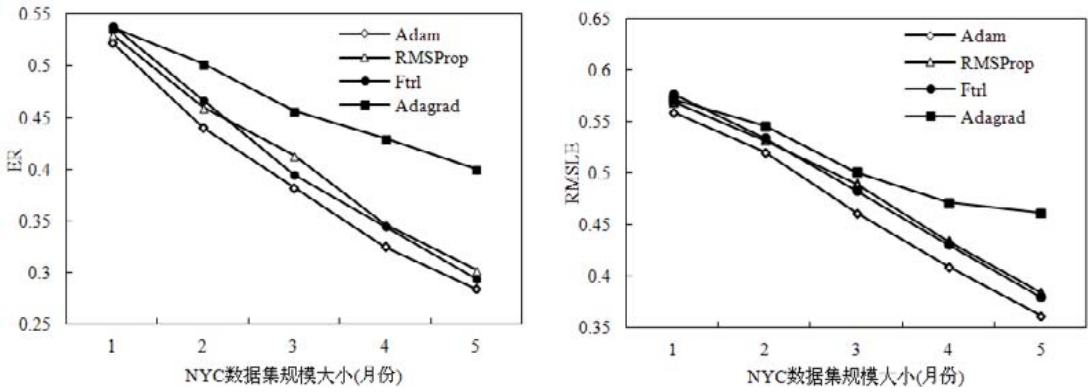
为了最小化损失函数, 本文分别使用流行优化算法 RMSProp[25], Ftrl[26], Adagrad[27]和 Adam[28]对模型进行训练. 图 22 和图 23 表示在不同数据规模下, 使用 4 种优化算法时, 借/还车数量预测性能的对比结果.

实验结果表明: 随着数据规模增大, 当使用 5 个月的数据进行训练时, 选择 Adam 算法优化模型得到的预测准确度最高.

图 22 中: 对于 *ER* 评估指标, 预测模型使用 Adam 算法相较于 RMSProp 算法提高了 5.1%, 相较于 Adagrad 算法提高了 19.1%, 相较于 Ftrl 算法提高了 4.5%; 对于 RMSLE 评估指标, 预测模型使用 Adam 算法相较于 RMSProp 算法提高了 4.1%, 相较于 Adagrad 算法提高了 10.3%, 相较于 Ftrl 算法提高了 3.9%.

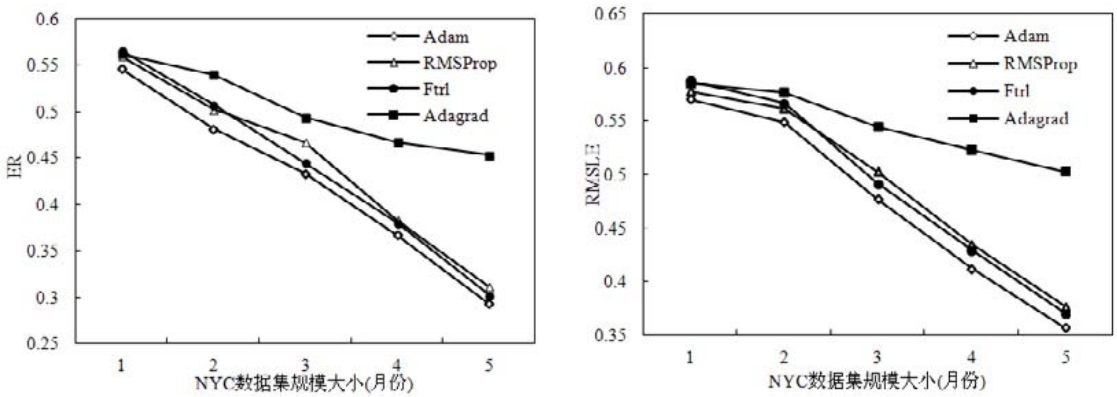
图 23 中: 对于 *ER* 评估指标, 预测模型使用 Adam 算法相较于 RMSProp 算法提高了 4.8%, 相较于 Adagrad 算法提高了 18.7%, 相较于 Ftrl 算法提高了 3.8%; 对于 RMSLE 评估指标, 预测模型使用 Adam 算法相较于 RMSProp 算法提高了 3.8%, 相较于 Adagrad 算法提高了 15.6%, 相较于 Ftrl 算法提高了 3.5%.

通过本节实验, 使用 Adam 算法优化预测模型效果最佳. 原因在于: Adam 算法可以在不同参数条件下自适应选择学习率, 使得更新所有参数权重的学习率不再单一, 因此预测性能得以提升.



(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 22 优化算法对借车数量预测性能的影响

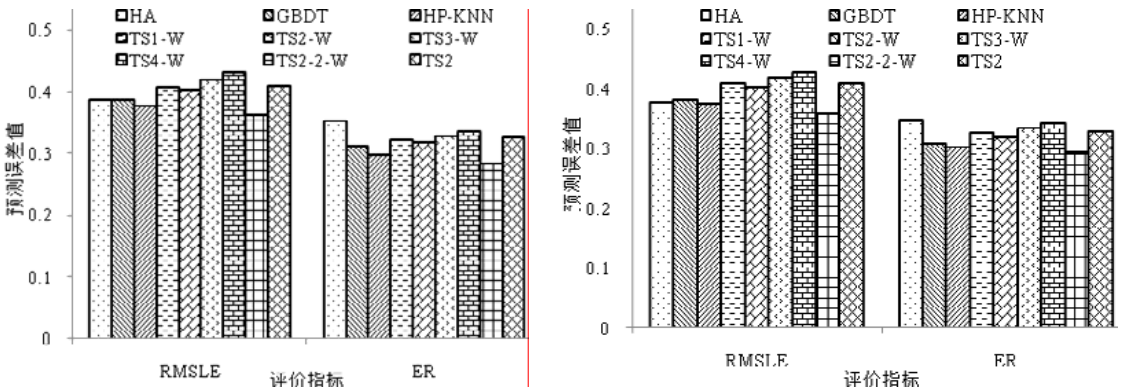


(a) 不同数据规模下 ER 值 (b) 不同数据规模下 RMSLE 值

图 23 优化算法对借车数量预测性能的影响

6.5.5 不同模型预测准确性对比

图 24 给出了本文所提预测模型 DeepML 同 HA, GBDT 和 HP-*K*NN 这 3 种模型的预测准确性对比结果.



(a) 不同模型借车数量预测性能 (b) 不同模型还车数量预测性能

图 24 不同模型预测结果对比

针对 DeepML 模型, 本文给出了 6 种关于 LSTM 层数和影响因素的组合变量的结果. 以 TS2-2-W 为例: 当前时间的前两个时间步(用 *TS*2 表示), 多特征 LSTM 网络有 2 个 LSTM 层(用 2 表示), 考虑天气因素(用 W 表

示). 注意: 其他 5 种 DeepML 模型中只使用了多特征 LSTM 网络的 1 个 LSTM 层, 所以没有用“2”表示.

通过实验结果可以发现: DeepML 模型对借车和还车数量的预测性能均表现为最佳. 主要原因在于: 单车需求量和天气都是随时间变化的, 多特征 LSTM 网络中记忆细胞具有多个门控结构, 使得 DeepML 模型可以长时间记忆过去几个时间段内单车系统的状态, 可以从时序上建立其与下一时刻单车需求的关系, 因此预测的准确性最高.

7 结束语

为了解决随时间变化共享单车系统不平衡问题, 本文提出一种新型基于站点聚类的共享单车需求预测模型. 根据历史行程记录和站点分布数据构建出单车转移网络, 得到每个站点的活跃度, 综合考虑站点位置和单车使用模式, 对站点进行二级聚类. 分析时间和天气因素对聚簇内单车需求量的影响, 选取关键特征构建二维向量, 使用多特征 LSTM 网络预测不同时间段聚簇内单车需求.

未来工作包括: (1) 考虑更多影响共享单车出行的因素, 比如突发事件, 进而预测极端情况下的单车需求; (2) 因为用户的用车行为是随机的, 需要融合用户相关信息, 如社交数据等, 分析用户个体的用车习惯, 进一步提高预测结果的准确性; (3) 由于目前解决站点内单车需求不平衡问题的方法是进行人工调度, 所以未来工作可以通过寻找站点间最优路径节省人工调度的成本.

References:

[1] http://www.hellobike.com

[2] Yang ZD, Hu J, Shu YC, *et al*. Mobility modeling and prediction in bike-sharing systems. In: Proc. of the 14th Annual Int'l Conf. on Mobile Systems, Applications and Services. New York: ACM, 2016. 165−178. [doi: 10.1145/2906338.2906408]

[3] Huang F, Qiao SJ, Peng J, *et al*. A bimodal gaussian inhomogeneous poisson algorithm for bike number prediction in bike-sharing system. IEEE Trans. on Intelligent Transportation Systems, 2019, 20(8): 2848−2857. [doi: 10.1109/TITS.2018.2868483]

[4] Ashqar HI, Elhenawy M, Almannaa MH, *et al*. Modeling bike availability in a bike-sharing system using machine learning. In: Proc. of the 5th IEEE Int'l Conf. on MT-ITS. Washington: IEEE, 2017. 374−378. [doi: 10.1109/MTITS.2017.8005700]

[5] Liu JJ, Sun LL, Chen WW, *et al*. Rebalancing bike sharing systems: A multi-source data smart optimization. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2016. 1005−1014. [doi: 10.1145/2939672.2939776]

[6] Fricker C, Gast N. Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity. Euro Journal on Transportation and Logistics, 2016, 5(3): 261−291. [doi: 10.1007/s13676-014-0053-5]

[7] Julia P, Przemyslaw AG, Ryota K, *et al*. Predicting the success of online petitions leveraging multidimensional time-series. In: Proc. of the 26th Int'l Conf. on World Wide Web. New York: ACM, 2017. 755−764. [doi: 10.1145/3038912.3052705]

[8] Qiao SJ, Han N, Wang JF, *et al*. Predicting long-term trajectories of connected vehicles via prefix-projection technique. IEEE Trans. on Intelligent Transportation Systems, 2018, 19(7): 2305−2315. [doi: 10.1109/TITS.2017.2750075]

[9] Li YX, Zheng Y, Zhang HC, *et al*. Traffic prediction in a bike-sharing system. In: Proc. of the 23rd SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM, 2015. 33:1−33:10. [doi: 10.1145/2820783.2820837]

[10] Chen LB, Zhang DQ, Wang LY, *et al*. Dynamic cluster-based over-demand prediction in bike sharing systems. In: Proc. of the 2016 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing. New York: ACM, 2016. 841−852. [doi: 10.1145/2971648.2971652]

[11] Feng SJ, Chen H, Du C, *et al*. A hierarchical demand prediction method with station clustering for bike sharing system. In: Proc. of the 3rd IEEE Int'l Conf. on Data Science in Cyberspace. Washington: IEEE, 2018. 829−836. [doi: 10.1109/DSC.2018.00133]

[12] Zhang XK, Fei S, Song C, *et al*. Label propagation algorithm based on local cycles for community detection. Int'l Journal of Modern Physics B, 2015, 29(5): 1550029. [doi: 10.1142/S0217979215500290]

[13] Schuijbroek J, Hampshire RC, Van Hoeve WJ. Inventory rebalancing and vehicle routing in bike sharing systems. European Journal of Operational Research, 2017, 257(3): 992−1004. [doi: 10.1016/j.ejor.2016.08.029]

[14] Lin L, He ZB, Peeta S. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation Research Part C: Emerging Technologies, 2018, 97: 258−276. [doi: 10.1016/j.trc.2018.10.011]

[15] Chai D, Wang LY, Yang Q. Bike flow prediction with multi-graph convolutional networks. In: Proc. of the 26th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM, 2018. 397−400. [doi: 10.1145/3274895.3274896]

[16] Lv YS, Duan YJ, Kang WW, *et al*. Traffic flow prediction with big data: A deep learning approach. IEEE Trans. on Intelligent Transportation Systems, 2015, 16(2): 865−873. [doi: 10.1109/TITS.2014.2345663]

[17] Xu YX, Wu WG, Wang SM, *et al*. Data center temperature prediction algorithm based on long short-term memory network. Computer Technology and Development, 2019, 29(12): 1−7 (in Chinese with English abstract). [doi: 10.3969/j.issn.1673-629X.2019.12.001]

[18] Su M, Wu C, Huang K, *et al*. Cell-coupled long short-term memory with *l*-skip fusion mechanism for mood disorder detection through elicited audiovisual features. IEEE Trans. on Neural Networks and Learning Systerms, 2020, 31(1): 124−135. [doi: 10.1109/TNNLS.2019.2899884]

[19] Cohen J, Cohen P, West SG, *et al*. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: Psychology Press, 2013. 379−384. [doi: 10.4324/9781410606266]

[20] Bai Z, Huang L, Chen JN, *et al*. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018, 29(4): 1029−1038 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[21] Zhou XL, Chen YG. Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in Chicago. Plos One, 2015, 10(10): e0137922. [doi: 10.1371/journal.pone.0137922]

[22] Chen LB, Ma XJ, Nguyen TMT, *et al*. Understanding bike trip patterns leveraging bike sharing system open data. Frontiers of Computer Science, 2017, 11(1): 38−48. [doi: 10.1007/s11704-016-6006-4]

[23] Chardon CMD, Caruso G, Thomas I. Bike-share rebalancing strategies, patterns, and purpose. Journal of Transport Geography, 2016, 55: 22−39. [doi: 10.1016/j.jtrangeo.2016.07.003]

[24] Caulfield B, O'Mahony M, Brazil W, *et al*. Examining usage patterns of a bike-sharing scheme in a medium sized city. Transportation Research Part A: Policy and Practice, 2017, 100: 152−161. [doi: 10.1016/j.tra.2017.04.023]

[25] Taqi AM, Awad A, Al-Azzo F, *et al*. The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance. In: Proc. of the 1st IEEE Conf. on Multimedia Information Processing and Retrieval. Washington: IEEE, 2018. 140−145. [doi: 10.1109/MIPR.2018.00032]

[26] Mcmahan HB, Holt G, Sculley D, *et al*. Ad click prediction: A view from the trenches. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2013. 1222−1230. [doi: 10.1145/2487575.2488200]

[27] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011, 12: 2121−2159. [doi: 10.1109/TNN.2011.2146788]

[28] Kingma D, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR 2015). San Diego, 2015. 1−15.

[29] Cai L, Gu J, Ma JH, *et al*. Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees. Energies, 2019, 12(1): 1−19. [doi: 10.3390/en12010159]

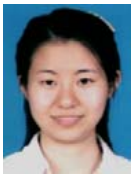
[30] Bhatia N, Vandana. Survey of nearest neighbor techniques. Int'l Journal of Computer Science and Information Security, 2010, 8(2): 302−305.

附中文参考文献:

[17] 徐一轩, 伍卫国, 王思敏, 等. 基于长短期记忆网络(LSTM)的数据中心温度预测算法. 计算机技术与发展, 2019, 29(12): 1−7. [doi: 10.3969/j.issn.1673-629X.2019.12.001]

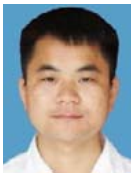
[20] 白琮, 黄玲, 陈佳楠, 等. 面向大规模图像分类的深度卷积神经网络优化. 软件学报, 2018, 29(4): 1029−1038. http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

乔少杰(1981－), 男, 博士, 教授, CCF 杰出会员, 主要研究领域为机器学习, 城市计算, 深度学习.



韩楠(1984－), 女, 博士, 副教授, 主要研究领域为机器学习.

岳昆(1979－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为人工智能.



易玉根(1986－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为机器学习, 深度学习.

黄发良(1975－), 男, 博士, 副教授, 主要研究领域为数据挖掘.

元昌安(1964－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为数据库.

丁鹏(1993－), 男, 硕士生, 主要研究领域为机器学习.

Gutierrez LA(1980 －), 男 , 博 士 , Researcher, 主要研究领域为机器学习.